粗集理论在污水参数软测量中的应用研究*

卿晓霞[1] 龙腾锐[1] 王 波[2] 余建平[1]
[1](重庆大学三峡库区生态环境教育部重点实验室 重庆 400045)
[2](重庆大学计算机学院 重庆 400044)

摘要 用粗糙集理论约简属性,消除冗余信息后建立了污泥体积指数的神经网络软测量模型。用某城市污水厂实际水质参数进行仿真实验。仿真结果表明,与未采用粗糙集进行预处理的模型相比,应用该模型不仅测量值的误差更小,而且输入参数从 9 个降至 4 个,大大降低了输入数据的维数,减少了神经网络的训练时间及训练步数,有利于软测量模型的实用化。

关键词 粗糙集 人工神经网络 软测量 污泥体积指数
中图分类号 X703.1 **文献标识码** A **国家标准学科分类代码** 560.5510

Application research of rough set theory in wastewater parameters soft measure

Qing Xiaoxia[1] Long Tengrui[1] Wang Bo[2] Yu Jianping[1]
[1](*Key Laboratory of the Three Gorges Reservoir Region's Eco-Environment, Ministry of Education, Chongqing University, Chongqing 400045, China*)
[2](*Computer College, Chongqing University, Chongqing 400044, China*)

Abstract SVI artificial neural network soft measurement model was established using rough set theory to reduce attributes and eliminate superfluous data. A simulation was carried out using data from a wastewater treatment plant. Simulation result indicates that compared with the model that does not use rough set theory for pre-processing, the proposed model has lower measurement error and reduces the number of input parameters from 9 to 4. The dimension of the input data is reduced greatly, the training time and steps of the artificial neural network are also reduced, which is an advantage for the soft measurement model to be used in practice.

Key words rough set artificial neural network soft sensor SVI

1 引 言

在活性污泥法污水处理工艺中,污泥体积指数(SVI)是评定活性污泥质量的重要指标之一,它反映了活性污泥的絮凝、沉降性能,并能及早发现污泥膨胀等异常现象。SVI 值过低,说明泥粒细小,无机物含量高,缺乏活性;过高,说明污泥沉降性能不好,并且已有产生污泥膨胀的可能[1]。然而,SVI 不能在线实时测量,在实际应用中都是通过人工测量、计算得到,一次测量耗时至少 2h 以上。如能实时检测 SVI,对于污水处理系统的运行管理、提高出水质量及防止污泥膨胀现象的发生都具有重要的作用。

基于人工神经网络(artificial neural network,ANN)的软测量技术是解决这类难测参数实时在线测量的有效方法。所谓软测量,就是根据某种最优准则,选择一组既与主导变量(难测变量)有密切联系又容易测量的变量(称辅助变量),通过构造某种数学关系,用计算机软件实现对主导变量的估计[2]。虽然 ANN 在自学习、并行处理、联想记忆、容错性及极强的非线性映射能力等方面具有突出的优越性,但由于它不能确定哪些知识是冗余的,哪些知识是有用的,易造成训练时间过于漫长,甚至难以收敛。ANN 的这个固有缺点是制约其软测量模型进一步实用化的主要因素之一。粗糙集(rough set)理论[3]是

* 本文于 2005 年 10 月收到,系重庆市自然科学基金(CSTC,2005BB7250)、国家"十五"科技攻关(2004BA604A01)资助项目。

用来研究不完整数据,不精确知识的表示、学习、归纳的方法,它的一个突出优点是具有很强的定性分析能力,即不需要预先给定某些特征或属性的数量描述,而是直接从给定问题的描述集合出发,通过不可分辨关系和不可分辨类确定问题的近似域,找出问题中的内在规律[4-5],通过简单的决策表定义条件属性和决策属性间的依赖关系,即输入空间与输出空间的映射关系,通过知识约简去掉冗余属性,可大大简化知识表达空间维数。

本文融合粗糙集理论和 ANN 两者的优势,把粗糙集作为 ANN 的前置系统,利用粗糙集对数据进行约简,在保留重要信息的前提下消除冗余的数据,简化网络结构,从而缩小神经网络的搜索空间,以提高训练速度。利用这种方法以某城市污水处理厂的实际污水水质参数数据为例,给出了基于粗糙集-神经网络的污水处理难测参数软测量的具体过程,并和未经粗糙集预处理的 ANN 模型进行比较,证明了该方法的有效性。

2 粗糙集理论概述[3,5-6]

粗糙集理论以不可分辨关系划分所研究论域的知识,形成知识表达系统;利用上、下逼近描述对象;通过知识约简,获得最简知识。

2.1 知识、不可分辨关系与基本集

在粗糙集理论中,知识被认为是一种分类能力。人们的行为是基于分辨现实的或抽象的对象的能力。

设 U 是非空有限论域,R 是 U 中的等效关系。二元对 $K=(U,R)$ 构成一个近似空间。$\forall(x,y)\in U\times U$,若 $(x,y)\in R$,则称对象 x 与 y 在近似空间 K 中是不可分辨的,不可分辨关系也被称为等效关系。U/R 表示 U 中对象由 R 构成的等效关系类族,它构成了 U 的一个划分。可以证明,U 上划分可以与 U 上的二元等效关系之间建立一一对应。U/R 的集合称为基本集,即为论域中相互间不可分辨的对象组成的集合,是组成论域知识的颗粒。

不可分辨关系概念是粗糙集理论的基石,它深刻地揭示出知识的颗粒状结构,是定义其他概念的基础。

2.2 粗糙集的上、下逼近和边界区

对于论域 U 上任意一个子集 X,X 不一定能用知识精确地描述,这时就用 X 关于 K 的一对下逼近和上逼近来"近似"地描述。其定义如下:

设 $[x]_R$ 表示所有与 x 不可分辨的对象所组成的集合,即由 x 决定的等效关系类。

集合 X 关于 R 的下逼近定义为:

$$\underline{R}(X)=\{x\in U:[x]_R\subseteq X\}。$$

$\underline{R}(X)$ 实际上是由那些根据已有知识判断肯定属于 X 的对象所组成的最大的集合,也称为 X 的正区,记作 $\mathrm{POS}(X)$。由根据已有知识判断肯定不属于 X 的对象组成的集合称为 X 的负区,记作 $\mathrm{NEG}(X)$。

集合 X 关于 R 的上逼近定义为:

$$\overline{R}(X)=\{x\in U:[x]_R\cap X\neq\phi\}。$$

$\overline{R}(X)$ 是由所有与 X 相交非空的等效类 $[x]_R$ 的并集,是那些可能属于 X 的对象组成的最小集合。显然,$\overline{R}(X)+\mathrm{NEG}(X)=$论域 U。

集合 X 的边界区定义为:$BN(X)=\overline{R}(X)-\underline{R}(X)$。

若 $BN(X)$ 是空集,则称 X 关于 R 是清晰的;反之则称集合 X 为关于 R 的粗糙集。

2.3 知识表示系统

粗糙集中的知识表达系统可表示为如下的四元有序组:

$$K=(U,A,V,\rho)$$

式中:U 为对象的集合,即论域;A 为属性的集合。$C\cup D=A$,$C\cap D=\phi$,C 和 D 分别为条件属性集和决策属性集;V 为属性值的集合,$V=\bigcup_{a\in A}V_a$,V_a 是属性的值域;ρ 为信息函数,$\rho_x:A\to V$,$x\in U$,确定了 U 中每一对象的属性值。

这种描述方式使知识表达系统可以用二维表格来表示,这样的表格称为决策表,它包含了相应知识表中所有范畴的描述和数据推导出的所有可能的规律。

2.4 知识约简与核

用决策表表达知识时,在决策表中经常会遇到一些多余的数据(知识),即从中删除一些数据而依然保持决策表的基本性质。这些被删除的数据则是这个信息表中的冗余数据。删除冗余数据的过程称为属性的约简。约简定义如下:

如果 C' 是一个满足如下条件 $\mathrm{IND}(C,D)=\mathrm{IND}(C',D)$ 的 C 中最小子集,则定义 $C'\subseteq C$ 为 C 一个关于 D 的约条件属性集。一个决策表可能含有多个约简,所有 D- 约简的交集称为关于 D 的核(D-core)。因为核是所有约简的交集,因而它包含于所有约简之中。核是属性中最重要的子集,删除任一核元素将会影响属性分类能力。

约简和核这 2 个概念很重要,是粗糙集理论的精华。粗糙集理论提供了搜索约简和核的方法。

3 粗糙集神经网络污水参数软测量方法

下面以某污水处理厂的 17 组实际运行数据为例,给出基于粗糙集-ANN 的 SVI 软测量具体实现方法:

(1) 初始决策表的形成。在获取运行数据后,首先以拉依达(Lauta)准则剔除数据中的粗大误差,再以多元线性回归的方法填补去除的数据,获得初始决策表如表 1 所示。然后将数据分成 2 部分,1～12 组数据用于训练

表 1　某污水厂实测运行数

COD/ N	TN/ TP	DO(po) mg·L⁻¹	t(po)/ ℃	F/ M(kgCOD/ (kgMLSS.d))	NH4(po) mg·L⁻¹	turbid (°)NTU	ORP(po) mV	DO mg·L⁻¹	SVI(po) mL·g⁻¹
4.53	10.54	2.1	16.4	0.360	28.86	118.2	318	1.6	248
6.90	8.99	1.9	18.8	0.764	49.98	150.0	253	1.1	280
6.39	9.60	3.5	17.4	0.190	34.86	145.8	388	1.1	220
5.28	10.04	0.6	17.4	0.389	32.86	156.5	336	1.3	377
7.20	8.67	3.1	18.3	0.189	37.86	177.2	403	0.9	240
8.18	8.45	3.2	18.8	0.262	39.86	196.5	373	0.6	250
7.38	8.70	4.9	19.2	0.153	46.98	176.0	242	1.2	241
4.59	14.73	3.2	19.8	0.254	53.10	175.2	324	0.9	221
6.36	11.76	1.5	19.2	0.562	52.10	120.9	209	1.0	264
4.62	10.81	0.6	18.0	0.042	21.12	88.6	290	2.5	307
4.94	12.75	1.9	16.9	0.198	18.62	69.1	315	2.6	293
4.95	16.74	1.7	17.3	0.305	26.61	77.1	266	1.3	247
6.83	9.24	4.4	17.8	0.108	51.98	191.2	287	1.3	248
7.52	9.65	1.9	18.8	0.654	47.36	135.6	302	1.3	282
6.90	8.99	1.9	18.8	0.726	49.98	150.0	253	1.1	277
6.43	11.07	2.5	19.1	0.323	46.23	146.8	232	1.2	238
7.31	8.53	1.4	19.0	0.648	48.61	156.5	217	0.9	249

注:po 表示曝气池内参数,其余表示进水参数。

网络,13～17 组数据用于测试网络。

(2) 属性值离散化。粗糙集方法是一类符号化分析方法,需要将连续的属性离散化。连续属性的离散就是将连续属性值域划分为若干个区间,每个区间用不同代码表示属性值。常用的方法有神经网络法、模糊聚类法、等频离散法、等宽离散法及利用相关领域知识的方法等。本文采用等频离散算法并结合相关领域知识,将各属性参数作离散化处理。

(3) 形成决策表。采用离散化后的属性值形成一张二维表格,每一行描述一个对象,每一列描述对象的一种条件属性,最后一列为决策属性。

(4) 属性约简。删除重复的对象及其条件属性,考察决策表的兼容性。如果删除该属性后决策表兼容,删除该属性,直到决策表最简单为止,最后得到信息表的多个约简如图 1 所示。

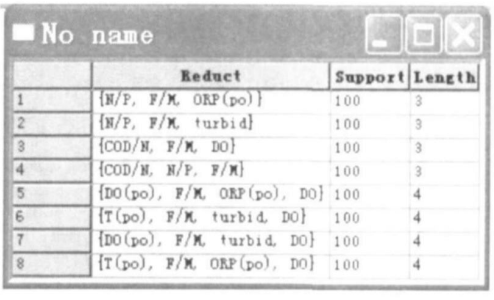

图 1　约简属性集

(5) 考虑 SVI 的影响因素以及各属性与 SVI 的相关系数,选择第 7 组为最佳约简属性集(即:DO(po),F/ M,turbid,DO),并对其进行规则提取,获取规则集如图 2 所示。

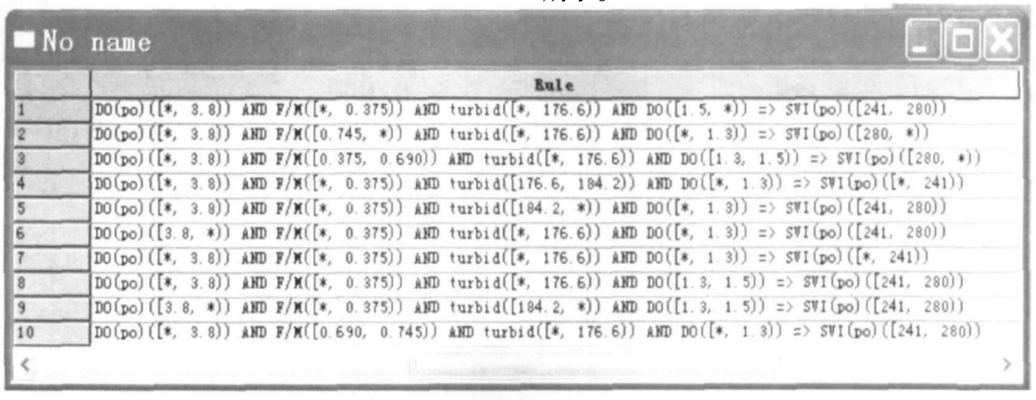

图 2　规则集

(6) 将上述规则集作为神经网络训练数据,网络结构为 4-7-1,如图 3 所示。用 BP 算法对其进行训练,再以剩下的 5 组数据对训练好的网络进行测试,结果如图 4 所示。

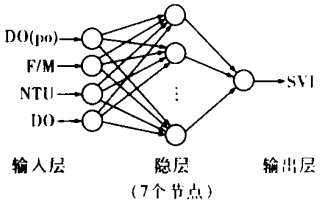

图 3 BP 网络结构图

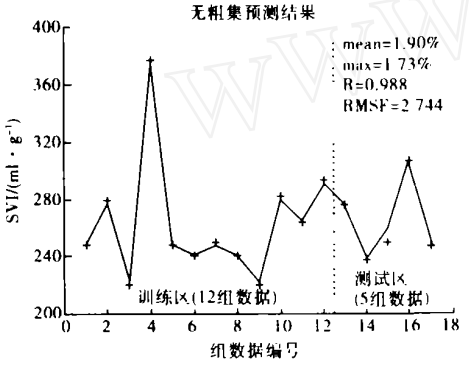

图 4 RS-BP 网络测量结果

(7) 再以原始训练数组直接输入神经网络,与上述结果进行对比。网络结构为 9-7-1,同样用 BP 算法对其进行训练,以测试数组进行测试,结果如图 5 所示。

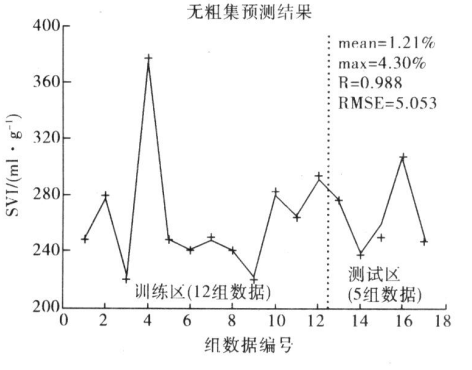

图 5 BP 网络测试结果

(8) 粗集 - BP 神经网络与 BP 神经网络测量结果对比如表 2 所示。

从表 2 可以看出,粗集 - BP 神经网络模型与 BP 神经网络模型相比,由于大大降低了输入变量的维数、简化了网络结构,不仅缩短了训练时间、减少了训练步数,而且由于属性约简求核,消除了样本中噪声数据的干扰,使测量结果的绝对误差平均值、相对误差平均值及均方根误差等都明显减少。

表 2 RS-BP 网络与 BP 网络测量结果对比

	BP 神经网络		粗集 BP 神经网络	
	绝对误差	相对误差(%)	绝对误差	相对误差(%)
平均值	3.12	1.21	2.33	0.90
最大值	10.71	4.30	4.30	1.73
训练步数	114		81	
训练时间/s	2.093		1.452	
RMSE	5.053		2.744	

4 结束语

本文通过实例说明了粗糙集理论在污水参数软测量神经网络模型预处理中的有效性。在 17 组污水处理厂水质参数中,原每组数据均含 10 个参数,其中 9 个输入参数,1 个输出参数。通过粗糙集理论对属性约简求核后,不仅输入参数的维数降到了 4 个,且获得了较好的预测效果。由于 SVI 基于粗集 - BP 神经网络软测量模型的 4 个输入参数均是可实时在线检测的易测变量,因此只要经过实际数据的训练,该模型用于 SVI 的实时在线检测现实可行。同时,本文提出的基于粗糙集-神经网络的软测量方法,对于污水处理系统中其他难测的重要水质参数(如硝酸氮 NO_3^- - N、生化需氧量 BOD_5、总氮 TN、总磷 TP 等)的实时在线测量具有参考价值。

参考文献

[1] 张自杰. 排水工程[M]. 北京:中国建筑工业出版社,1996.

[2] 于静江,周春晖. 过程控制中的软测量技术[J]. 控制理论与应用,1996, 13(4): 137-144.

[3] PAWLAK Z. Rough set[J]. International Journal of Information and Computer Science, 1982, 11(5): 341-356.

[4] SWINIARSKI R, HARGIS L. Rough sets as front end of neural-networks texture classifiers[J]. Neurocomputing, 2001, 36:85-102.

[5] 曾黄磷. 粗集理论及其应用[M]. 重庆:重庆大学出版社,1996

[6] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001.

作者简介

卿晓霞,女,副教授,硕士生导师,主要研究方向为水处理自动控制理论及技术。

E-mail:qxx118@126.com